IOP Conf. Series: Materials Science and Engineering **210** (2017) 012028 doi:10.1088/1757-899X/210/1/012028

Time Series Analysis and Forecasting of Wastewater Inflow into Bandar Tun Razak Sewage Treatment Plant in Selangor, Malaysia

Taher Abunama and Faridah Othman*

Department of Civil Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia.

E-mail: *faridahothman@um.edu.my

Abstract. Analysing the fluctuations of wastewater inflow rates in sewage treatment plants (STPs) is essential to guarantee a sufficient treatment of wastewater before discharging it to the environment. The main objectives of this study are to statistically analyze and forecast the wastewater inflow rates into the Bandar Tun Razak STP in Kuala Lumpur, Malaysia. A time series analysis of three years' weekly influent data (156weeks) has been conducted using the Auto-Regressive Integrated Moving Average (ARIMA) model. Various combinations of ARIMA orders (p, d, q) have been tried to select the most fitted model, which was utilized to forecast the wastewater inflow rates. The linear regression analysis was applied to testify the correlation between the observed and predicted influents. ARIMA (3, 1, 3) model was selected with the highest significance R-square and lowest normalized Bayesian Information Criterion (BIC) value, and accordingly the wastewater inflow rates were forecasted to additional 52weeks. The linear regression analysis between the observed and predicted values of the wastewater inflow rates showed a positive linear correlation with a coefficient of 0.831.

1. Introduction

Sewage treatment plants (STPs) are among the most valuable infrastructures in countries' development. The main objective of STP's is to treat the collected wastewater sufficiently to prevent negative impacts to human health, aquatic life, and the surrounding environment. STP's capacity and treatment processes must be designed and operated carefully in order to provide reliable treatment despite fluctuating characteristics, such as inflow and organic loading of the influent waste stream, in order to maintain compliance environmental permit limits and effluent standards.

For that, it is significant to evaluate and predict the design loads periodically (e.g. 3-5 years), because influent hydraulic and loadings parameters can vary considerably depending on the population that is being served, vacations and even tourist inflow can affect the inflow rate of wastewater.

Forecasting and simulating STP's inflow rates are valuable in order to define the average as well as peak flow rates, which assess in future planning of collection and treatment facilities. This can be conducted based on the previous observed and recorded inflow rate values at regular time intervals, through time series analysis of sewage inflow rates into treatment plants.

Box-Jenkins [1] or Autoregressive Integrated Moving Average (ARIMA) models are able to fulfil this task, and give an accurate prediction. ARIMA model consists of an integrated component (d), which performs differencing of the time series to make it stationary [2].

Another two components are autoregressive AR (p) and moving average MA (q); AR component correlates the relation between the current value and the past value of time series, while, MA captures the duration of random shock in the series.

These techniques have been well established and used for predicting hydro meteorological parameters in various studies [3]–[8].

In this study, the for ARIMA model has been applied for a time series of Bandar Tun Razak STP's sewage inflow data. And best fitting ARIMA model were selected and assessed using linear regression analysis between the observed and predicted values.

2. Methodology

2.1. Bandar Tun Razak (BTR) STP

Bandar Tun Razak STP is located on Jalan 11/118b, Desa Tun Razak, 56000 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia. The sewage plant is operated and managed by the Indah Water Konsortium Sdn Bhd (IWK). Total area of the plant is about ten acres, while the reserved area is six acres (Figure 1).

BTR STP built to serve part of Kuala Lumpur, with daily design capacity of 25,000 m3 and equivalent to 100,000 populations. Currently, the plant receives and treats about 11,700 m3 per day which equivalent to 52,000 populations. Sequential Batch Reactor (SBR) treatment system is equipped in the plant.

Figure 1. Plan view of BTR STP (Google earth).

2.2. Data collection

The wastewater inflow rates (Q) data was obtained from Bandar Tun Razak STP management, and it covered three continuous years on a weekly basis between 2011 and 2013, and the average inflow rate was about 16,711 m3. The weekly laboratory measurements of sewage inflow rate and loading parameters such as biological oxygen demand (BOD), chemical oxygen demand (COD), suspended solids (SS) and ammoniacal nitrogen (NH3-N), are authorized by the Department of Environment (DOE) to guarantee meeting the DOE standards of STP's.

2.3. Autoregressive Integrated Moving Average (ARIMA)

ARIMA or Box-Jenkins [12] model is considered the most popular and effective statistical models for time series forecasting. It based on generating a liner function extracted from the past observations of a time series in order to forecast the future values [9].

The linear function is consisting of three parametric components, Auto-Regression (AR), Integration (d) and Moving Average (MA) [1]. This can be illustrated in the form ARIMA (p, d, q).

In auto-regression (AR) or ARIMA (p, 0, 0) model, of order "p", the value of current output Zt (Observed value) depends upon the prior outputs "p" and the current inputs "et" (independent random shock). Therefore, the AR (p) equation can be written as:

$$Zt = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots. + \phi_p Z_{t-p} + e_t \quad (1)$$

While, in moving average (MA) or ARIMA (0, 0, q) model of order "q", the current output Zt (Observed Value) depends on the current input and prior inputs "q". MA (q) is represented as:

$$Zt = e_1 - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots. - \theta_q e_{t-q} \quad (2)$$

However, Autoregressive Moving Average Model (ARMA) of order (p, q) combines both AR and MA elements. An ARIMA (p, 0, q) or ARMA (p, q) is a model for a time series that depends on p past values of itself and on q past random terms e_t. This method has the form of:

$$Zt = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots. + \phi_p Z_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots. - \theta_q e_{t-q} \quad (3)$$

The Box-Jenkins models require a stationary time series data; therefore, a non-stationary data is always transformed to induce mean stationarity. A difference of order one leads to the subtraction of each observed value with the neighbouring value, which gives the new time series. Hence term "d" is referred TO the degree of ordinary differencing, applied to achieve series stationarity:

$$Yt = Z_t - Z_{t-1} \quad (4)$$

After applying the ARMA model to the differenced time series, the differencing transformation is reversed to reclaim the original values obtained by the modelled values and "integration" ("d" times) is done. A process in which the d^{th} order differencing is involved is called an Integrated process of order d, it is denoted by the notion I (d). A combination of AR, MA and I models is called an ARIMA (p, d, q) model of order (p, d, q).

2.4. Model Development

Time series analysis and forecasting of waste water inflow in Bandar Tun Razak sewage treatment plant (STP) were performed using a historical record of 3 years (2011-2013). Sewage inflow data was provided on weekly interval basis during the study period (156 weeks).

ARIMA model was applied in this study through the following steps: model identification and estimation, diagnostic checking and forecasting. The identification test is done to obtain the value of order of differencing 'd' in ARIMA (p, d, q) and also the values of AR and MA operators. The appropriate orders of the ARIMA (p, d, q) model are usually determined through the Box-Jenkins model building methodology [8]. IBM SPSS statistics 22 software was used in this study. In addition, linear regression analysis was used to compare between the observed and predicted values.

2.5. Model performance tests

In order to judge the modelling accuracy and select the most fitted ARIMA model configurations, different performance criterions such as R-square, stationary R-square, root mean square error (RMSE), mean absolute percentage error (MAPE), and normalized Bayesian Information Criterion (BIC) were used to select the best fitting.

3. Results and discussion

3.1. Model Identification and estimation

A sequence graph of sewage inflow data (156 weeks) in Bandar Tun Razak was plotted to check the stationarity of analysed data as shown in Figures 1. By computing the autocorrelation and partial autocorrelation coefficients (ACF and PACF), the data was found to be non-stationary as shown in Figures 2.

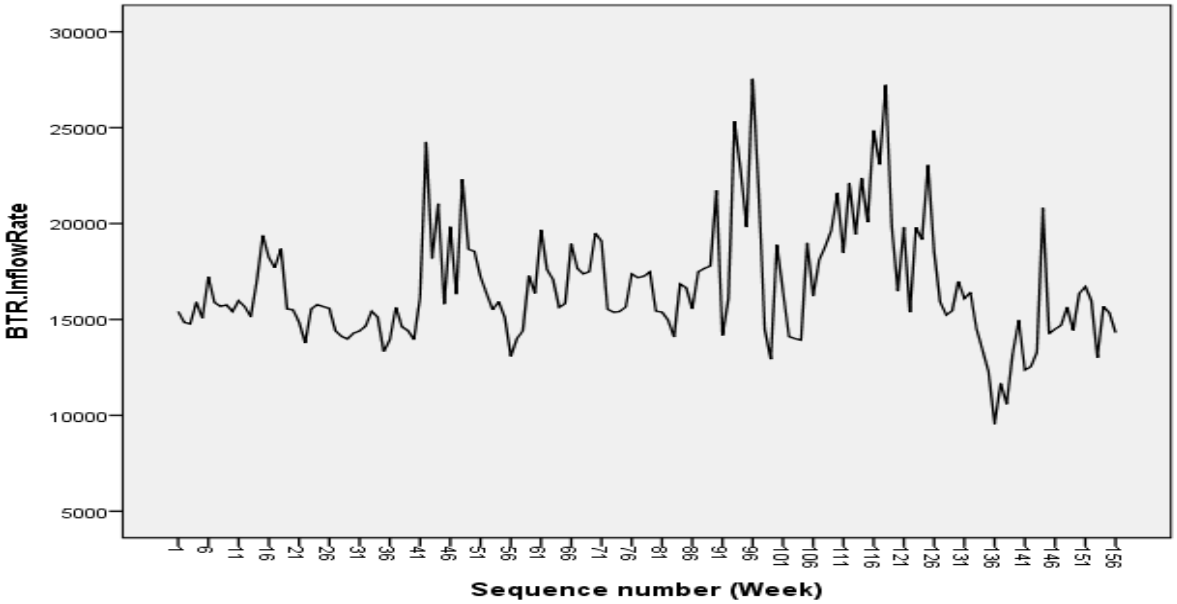

Figure 2. Sequence plot of the weekly inflow at BTR STP (152weeks).

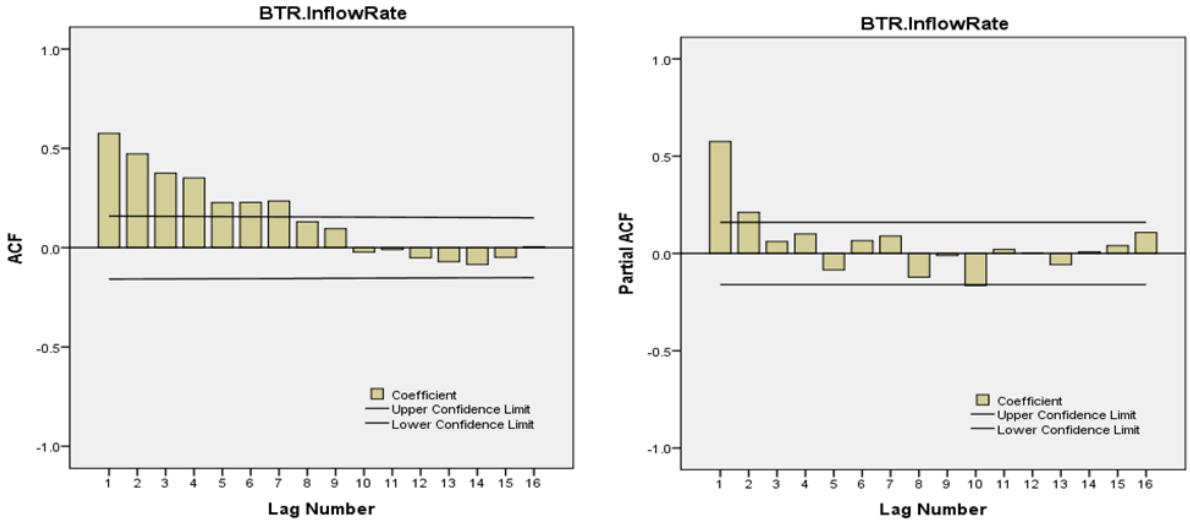

Figure 3. ACF and PACF of weekly inflow data.

Therefore, the first order differencing of the data series was applied (Figures 3). The obtained differenced data was tested for stationarity by ACF and PACF as shown in Figure 4. After examining (ACF and PACF) plots and its associated tables (Tables 7 in Appendix), it was concluded that the data become stationary in order to start applying ARIMA models.

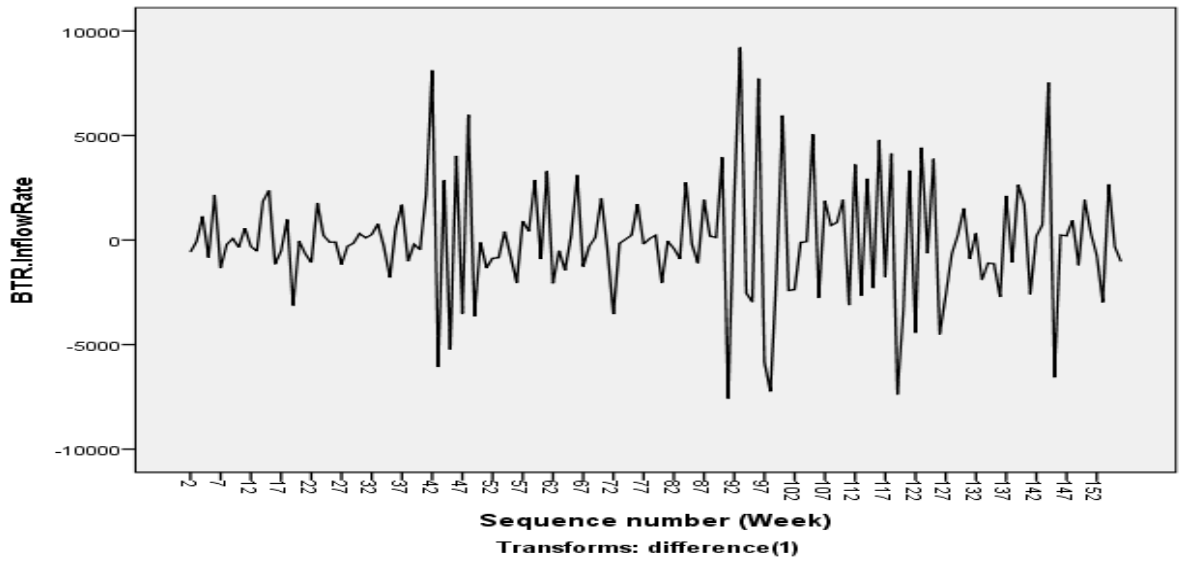

Figure 4. Sequence plot of the first order differencing of data.

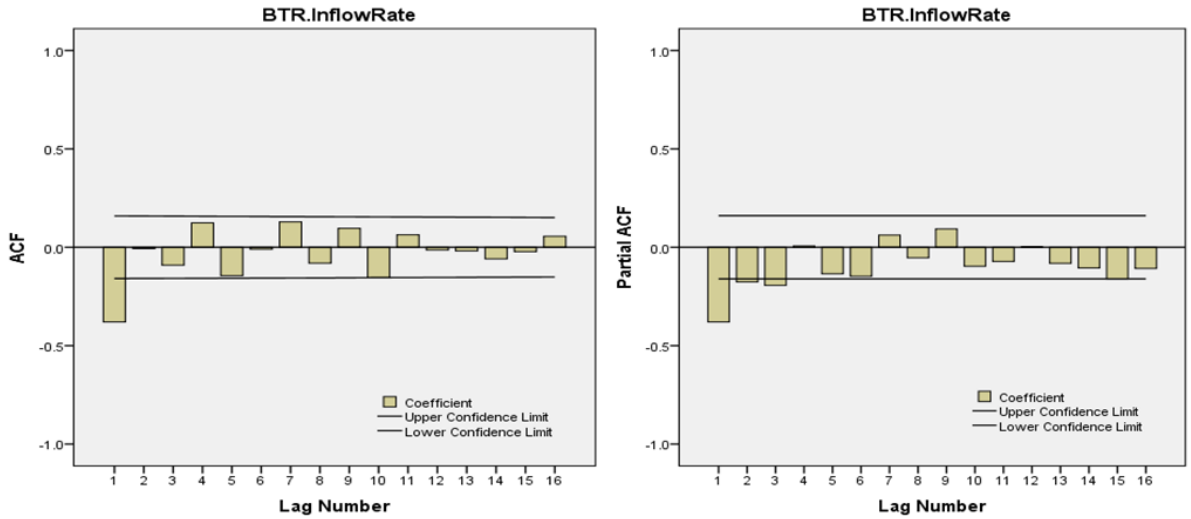

Figure 5. ACF and PACF of the first order differencing series.

Applying of ARIMA model was through trying various orders of 'p' and 'q' with the difference of one (d=1) to choose the best fitting ARIMA model. Among different configurations of ARIMA models orders, the best fitting model was chosen based on high stationary R-Square value, good value of R-Square and low values of RMSE, MAPE and Normalized BIC as illustrated in Table 1.

Table 1. Suggested ARIMA models.

Model (p, d, q)	Stationary R-Squared	R-Squared	RMSE	MAPE	Normalized BIC	Outliers
(0, 1, 1)	0.502	0.578	1990.0	8.79	15.42	5
(0, 1, 2)	0.525	0.598	1955.7	8.63	15.45	6
(0, 1, 3)	0.608	0.669	1788.8	8.08	15.34	7
(1, 1, 0)	0.573	0.639	1853.5	8.24	15.34	7
(1, 1, 1)	0.525	0.598	1955.7	8.64	15.45	6
(1, 1, 2)	0.505	0.581	1996.8	8.58	15.49	5
(1, 1, 3)	0.486	0.566	2041.0	8.93	15.57	5
(2, 1, 0)	0.507	0.583	1985.8	8.73	15.45	5
(2, 1, 1)	0.51	0.586	1986.4	8.50	15.48	5
(2, 1, 2)	0.486	0.566	2040.9	8.92	15.57	5
(2, 1, 3)	0.583	0.648	1850.8	8.18	15.44	6
(3, 1, 0)	0.593	0.656	1823.1	8.10	15.37	7
(3, 1, 1)	0.601	0.662	1812.4	8.01	15.40	7
(3, 1, 2)	0.511	0.586	1998.6	8.52	15.56	5
(3, 1, 3)	0.615	0.675	1791.3	8.01	15.44	7

The best suitable model for inflow rate of Bandar Tun Razak STP was found to be ARIMA (3, 1, 3). Main parameters of the selected model are given in the following tables:

Table 2. Statistics of the selected ARIMA (3, 1, 3) model.

Best Fit Model Statistics						Ljung-Box			Outliers
Stationary R-Squared	R-Squared	RMSE	MAPE	Max APE	Normalized BIC	Statistics	df	Sig.	7
0.615	0.675	1791.2	8.012	39.50	15.437	27.1934	12	0.007	

Table 3. Parameters of the selected ARIMA model

Parameters		Estimate	SE	t	Sig.
Constant		10.01	90.10	0.111	0.912
AR	Lag 1	-0.704	0.166	-4.228	0
	Lag 2	-0.694	0.183	-3.782	0
	Lag 3	-0.359	0.154	-2.328	0.021
Difference		1			
MA	Lag 1	-0.272	0.173	-1.574	0.118
	Lag 2	-0.650	0.154	-4.212	0
	Lag 3	0.234	0.158	1.483	0.140

Table 4. Outliers statistics

Outliers (Week No.)	Type	Estimate	SE	t	Sig.
42	Additive	5732.9	1257.7	4.558	0
90	Additive	6448.2	1405.3	4.589	0
93	Additive	12253.6	1727.7	7.093	0
94	Level Shift	7027.3	1949.0	3.606	0
98	Level Shift	-9826.9	1589.3	-6.183	0
118	Additive	6282.8	1217.3	5.161	0
144	Additive	7163.8	1233.3	5.809	0

3.2. Diagnostic checking

The selected model was tested and verified by examining the residuals ACF and PACF of various orders, which indicated a "good fit" of the model as shown in Figure 5. Autocorrelations up to 24 lags were evaluated and their significance was verified by Box-Ljung statistic as illustrated in Table 8 in Appendix.

Clearly, we can notice that almost all lags were within the reasonable level in residual ACF and residual PACF. Therefore, this refers that the selected ARIMA (3, 1, 3) model can be used for inflow rate analysis in Bandar Tun Razak STP.

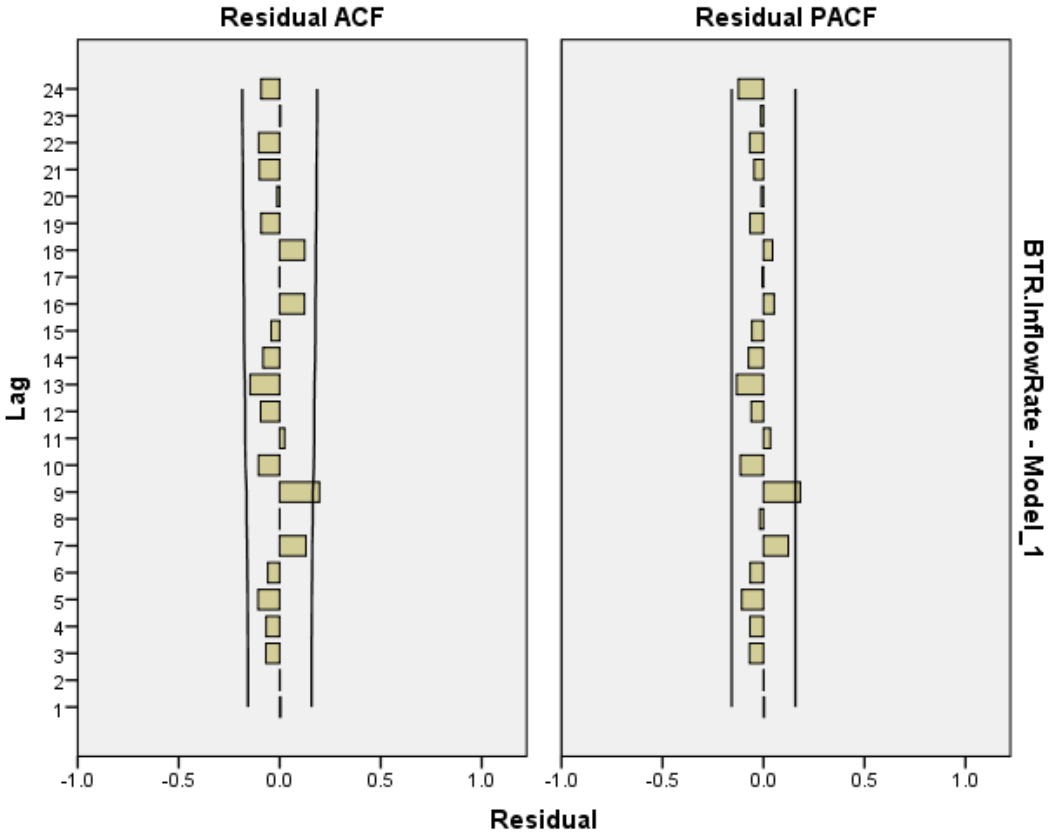

Figure 6. Residuals of ACF and PACF of the selected ARIMA model

The linear regression model was carried out between the observed inflow and predicted inflow values of ARIMA model. As shown in Table 5, the correlation coefficient of predicted inflow was 0.83, which suggests a good positive linear correlation.

Table 5. Linear Regression Model statistics.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df 1	df2	Sig. F Change	
1	0.83	0.691	0.689	1607.519	0.691	342.11	1	153	0	1.857

The Normal P-P Plot of Regression Standardized Residual showed a random scatter of the points with a constant variance without any outliers. Since the points are close to the diagonal line (Figure 6), it is understood that the residuals are approximately normally distributed.

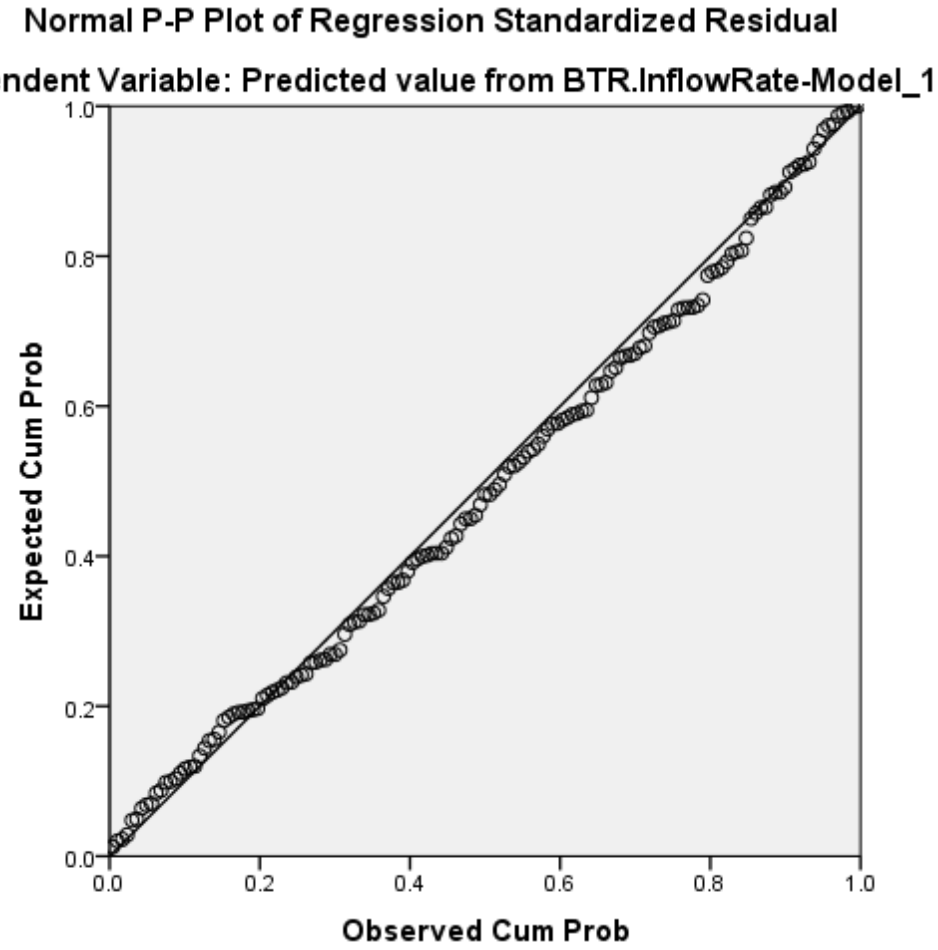

Figure 7. Normal P-P plot of regression standardized residual.

3.3. Forecasting

The best fitted ARIMA (3,1,3) was used to forecast the inflow rate till 208 weeks (4 years). The forecasted values are tabulated in Table 6, while the observed and predicted values with the confidential limits are shown in the Figure 7.

Table 6. Summary of Bandar Tun Razak STP inflow rate forecasting.

Period	Observed Inflow		Predicted Inflow		Lower Confidential Limit		Upper Confidential Limit	
Week 1-26	15,416	15,577	15,426	16,152	10,950	12,613	19,902	19,692
Week 27-52	14,413	16,352	14,809	18,604	11,270	15,066	18,348	22,141
Week 53-78	15,511	17,252	17,476	18,053	13,939	14,516	21,014	21,591
Week 79-104	17,479	13,926	16,935	15,174	13,398	11,636	20,473	18,711
Week 105-130	18,987	16,980	15,467	16,520	11,929	12,982	19,004	20,057
Week 131-156	16,087	14,314	15,724	16,578	12,186	13,041	19,261	20,116
Week 157-182	-	-	13,382	14,325	9,844	2,862	16,919	25,788
Week 183-208	-	-	14,332	14,582	2,667	(1,334)	25,998	30,498

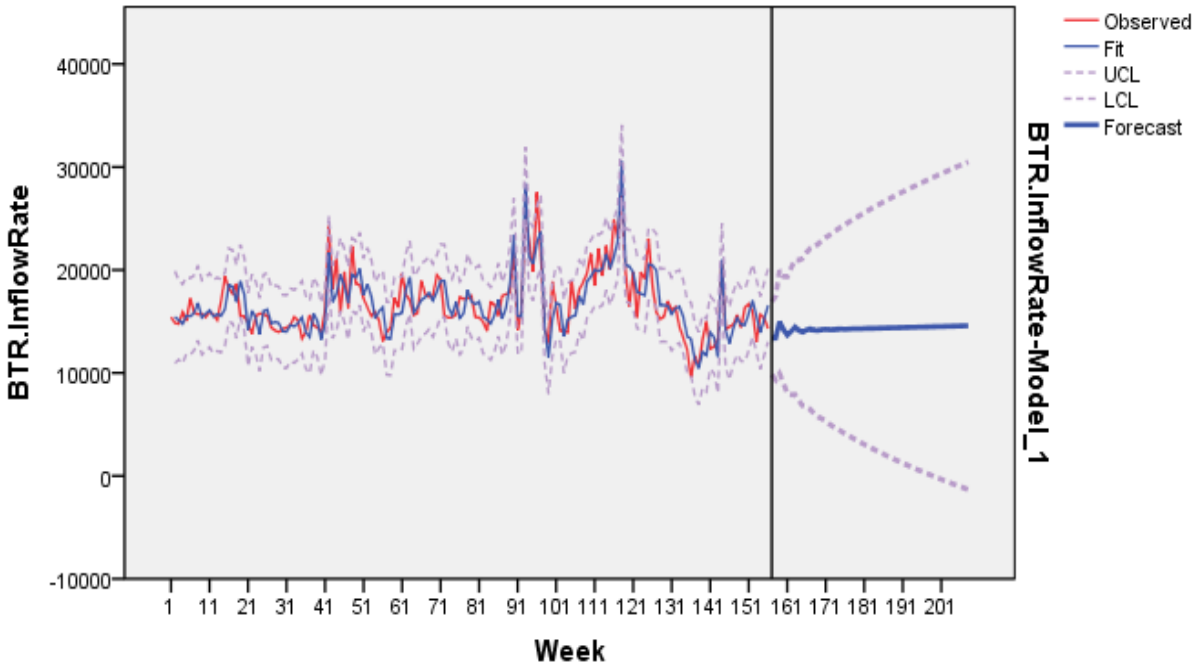

Figure 8. Forecasted ARIMA (3, 1, 3) model

4. Conclusions

In this study, time series ARIMA modelling of a weekly sewage inflow into one of the main STP's in Kuala Lumpur, Malaysia was successfully conducted. The three continues years (152weeks) data collected by the Bandar Tun Razak (BTR) STP's management was found to be non-stationary, thus it was transformed to the first order differencing (d=1) to make it stationary. Fifteen ARIMA models of various orders of 'p' and 'q' were applied on to the transformed data to select the best fitted model. Based on the diagnostics like high R^2 value and low normalized Bayesian Information Criterion (BIC), the ARIMA (3, 1, 3) was found to be the best fitted model. The linear regression model was applied between the observed and predicted values, and it showed a positive linear correlation with a correlation coefficient of 0.83. By this linear regression analysis, it was understood that there was no much variation between the observed and predicted data. The best fitted ARIMA (3, 1, 3) model forecasted the inflow

rates to additional one year (52weeks); therefore, this study can be considered for future design planning of the BTR STP.

Acknowledgment
The authors would like to acknowledge the University of Malaya Research Grant (FL001-13SUS) and Fundamental research Grant by Ministry of Higher Education (FP016-2014A) for the financial support. We are most grateful and would like to thank the reviewers for their valuable suggestions, which have led to substantial improvement of the article.

References
[1] G. E. P. Box and G. M. Jenkins 1976 *Time series analysis : forecasting and control*. Holden-Day
[2] A. Pankratz, 2009 *Forecasting with univariate Box-Jenkins models: Concepts and cases*.
[3] H. Yan and Z. Zou, 2013 *Journal of Convergence Information Technology* **8** 59–70.
[4] M. Valipour, M. E. Banihabib, and S. M. R. Behbahani 2013 *Journal of Hydrology* **476** 433–441.
[5] A. O. Pektaş and H. Kerem Cigizoglu 2013 *Journal of Hydrology* **500** 21–36.
[6] P. Narayanan, A. Basistha, S. Sarkar, and S. Kamna 2013 *Comptes Rendus Geoscience* **345** 22–27.
[7] J. R. Kim, J. H. Ko, J. H. Im, S. H. Lee, S. H. Kim, C. W. Kim, an T. J. Park 2006 *Water Science and Technology* **53** 185–19.
[8] G. P. Zhang 2003 *Neurocomputing* **50** 159–175.
[9] K. Yürekli, H. Simsek, B. Cemek, and S. Karaman 2007 *Building Environment*.

Appendix

- ACF and PACF statistics of the first order differencing series:

Autocorrelation			Box-Ljung Statistic			Partial Autocorrelation		
Lag	Value	Std. Error	Value	df	Sig.	Lag	Value	Std. Error
1	-0.380	0.080	22.77	1	0	1	-0.380	0.08
2	-0.007	0.079	22.77	2	0	2	-0.176	0.08
3	-0.091	0.079	24.09	3	0	3	-0.194	0.08
4	0.124	0.079	26.57	4	0	4	0.006	0.08
5	-0.145	0.079	29.96	5	0	5	-0.135	0.08
6	-0.010	0.078	29.98	6	0	6	-0.147	0.08
7	0.129	0.078	32.72	7	0	7	0.062	0.08
8	-0.081	0.078	33.81	8	0	8	-0.054	0.08
9	0.096	0.077	35.35	9	0	9	0.094	0.08
10	-0.154	0.077	39.33	10	0	10	-0.097	0.08
11	0.063	0.077	40.01	11	0	11	-0.072	0.08
12	-0.014	0.077	40.04	12	0	12	0.003	0.08
13	-0.018	0.076	40.10	13	0	13	-0.081	0.08
14	-0.059	0.076	40.70	14	0	14	-0.105	0.08
15	-0.022	0.076	40.78	15	0	15	-0.161	0.08
16	0.056	0.076	41.33	16	0	16	-0.108	0.08

- Residual of ACF and PACF of the selected ARIMA model:

Lag	ACF		PACF	
Lag 1	.005	.080	.005	.080
Lag 2	.002	.080	.002	.080
Lag 3	-.069	.080	-.069	.080
Lag 4	-.067	.081	-.067	.080
Lag 5	-.107	.081	-.107	.080
Lag 6	-.059	.082	-.065	.080
Lag 7	.130	.082	.122	.080
Lag 8	.000	.084	-.018	.080
Lag 9	.197	.084	.182	.080
Lag 10	-.105	.087	-.115	.080
Lag 11	.025	.087	.034	.080
Lag 12	-.095	.087	-.061	.080
Lag 13	-.147	.088	-.133	.080
Lag 14	-.083	.090	-.075	.080
Lag 15	-.042	.090	-.057	.080
Lag 16	.122	.090	.053	.080
Lag 17	-.001	.091	-.006	.080
Lag 18	.124	.091	.046	.080
Lag 19	-.092	.092	-.067	.080
Lag 20	-.013	.093	-.011	.080
Lag 21	-.101	.093	-.047	.080
Lag 22	-.104	.094	-.068	.080
Lag 23	.004	.094	-.013	.080
Lag 24	-.092	.094	-.123	.080

- Histogram chart of liner regression analysis: